PRISCILA NEVES FARIA

AVALIAÇÃO DE MÉTODOS PARA DETERMINAÇÃO DO NÚMERO ÓTIMO DE *CLUSTERS* EM ESTUDO DE DIVERGÊNCIA GENÉTICA ENTRE ACESSOS DE PIMENTA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2009

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
F224a
2009

Faria, Priscila Neves, 1984-
Avaliação de métodos para determinação do número ótimo de *clusters* em estudo de divergência genética entre acessos de pimenta / Priscila Neves Faria. – Viçosa, MG, 2009.
xi, 54f.: il. (algumas col.) ; 29cm

Inclui anexos.
Orientador: Paulo Roberto Cecon.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f. 40-43.

1. Estatística - Análise multivariada. 2. Estatística - Análise por agrupamento. 3. Melhoramento genético - Método estatístico - Programas de computador. 4. Biometria - Programas de computador. 5. GENES (programa de computador). I. Universidade Federal de Viçosa. II.Título.

CDD 22.ed. 519.535

PRISCILA NEVES FARIA

AVALIAÇÃO DE MÉTODOS PARA DETERMINAÇÃO DO NÚMERO ÓTIMO DE *CLUSTERS* EM ESTUDO DE DIVERGÊNCIA GENÉTICA ENTRE ACESSOS DE PIMENTA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 19 de janeiro de 2009.

Prof. Luiz Alexandre Peternelli

Prof. Cosme Damião Cruz
(Co-Orientador)

Prof. Antônio Policarpo S. Carneiro

Prof. Fernando Luiz Finger

Prof. Paulo Roberto Cecon
(Orientador)

“Deus tem consciência do teu limite. Ele só quer o que tu podes, porém se quiseres superar-te... ele te dará mais força (Walter Grando).”

A DEUS dedico esta vitória.

Aos meus queridos pais Mozart e Elena, exemplo de amor, confiança, apoio e luta na educação dos filhos.

Ao meu irmão Leandro, cujo apoio e incentivo foram importantes nesta etapa da minha vida profissional.

À minha avó Rosa Luzia Neves (*in memoriam),* fonte de orações, eterno carinho e amor.

À minha avó Rosa Maria Neves, fonte de orações e exemplo de força e superação.

À minha família, fortaleza em todos os momentos da minha vida.

Ao meu namorado Filipe, fonte de compreensão nos momentos difíceis, amizade, paciência, amor e carinho.

OFEREÇO

AGRADECIMENTOS

À Universidade Federal de Viçosa, por me acolher como estudante durante estes anos.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, pela concessão da bolsa de estudo.

A Deus, que me concedeu essa oportunidade e me deu forças para enfrentar os obstáculos. Obrigada, meu Deus, pela sabedoria, pelo amor e pela proteção. Obrigada pelas pessoas boas que colocou em meu caminho, pois elas me ajudaram a vencer e tornar minha caminhada mais tranqüila.

À minha família, fonte de amor inesgotável, que me incentivou e me carregou no colo nos momentos de fraqueza. Aos meus pais, Elena e Mozart, pelos sacrifícios feitos para que eu chegasse até aqui, por toda educação que recebi, pelo amor incondicional, pelo incentivo e por serem exemplo para mim; sem vocês, nada disso seria possível. Ao meu irmão, Leandro, pela torcida e pelo companheirismo concedidos ao longo de todo esse tempo. Obrigada pelo amor de vocês!

Ao Filipe, pela paciência, amor e carinho. Obrigada por tudo que fez e faz por mim e por nós, por me incentivar, me mostrando que sou capaz. Obrigada por fazer a minha vida mais feliz e por ser tão especial. Você também faz parte desta conquista.

Ao Prof. Paulo Roberto Cecon, por me orientar e acreditar em mim, mesmo nas horas em que nem eu acreditava. Obrigada por todos os ensinamentos e por toda paciência. Você, além de orientador, é também um amigo que sempre me mostrou a direção correta dos passos a serem tomados. Agradeço também aos meus co-orientadores, Professores Fabyano Fonseca e Silva e Cosme Damião Cruz, pelos ensinamentos durante o mestrado, pelos conselhos e sugestões neste trabalho. A contribuição de vocês no meu aprendizado foi valiosa.

A todos do corpo docente do Departamento de Informática, da Universidade Federal de Viçosa, pela formação e pelos conhecimentos recebidos.

Ao Altino Alves de Souza Filho, por sempre se prontificar a ajudar perante as burocracias do mestrado e por sempre me receber em sua sala com um sorriso no rosto e um abraço. Agradeço também a todos os funcionários, pela atenção e pela ajuda nos momentos que precisei.

A todos os colegas, do mestrado e da graduação, que me ajudaram e contribuíram direta ou indiretamente para a realização desta conquista.

BIOGRAFIA

PRISCILA NEVES FARIA, filha de Mozart Neves de Faria e Elena Maria de Faria, nasceu no dia 03 de fevereiro de 1984, em Belo Horizonte, MG.

Em 1988, mudou-se para a cidade de Divinópolis, MG, onde concluiu o Ensino Médio em 2001.

Em 2002, ingressou no Curso de Matemática da Universidade Federal de Viçosa (UFV), em Viçosa, MG, graduando-se em março de 2007.

Nesse mesmo mês e ano, ingressou no Programa de Pós-Graduação, em nível de Mestrado, em Estatística Aplicada e Biometria da UFV, submetendo-se à defesa da dissertação em janeiro de 2009.

SUMÁRIO

Página

RESUMO

FARIA, Priscila Neves, M.Sc., Universidade Federal de Viçosa, janeiro de 2009. **Avaliação de métodos para determinação do número ótimo de *clusters* em estudo de divergência genética entre acessos de pimenta.** Orientador: Paulo Roberto Cecon. Co-Orientadores: Fabyano Fonseca e Silva e Cosme Damião Cruz.

Muitas vezes, a interpretação dos resultados em análise de agrupamentos é feita de forma subjetiva, isto é, através da inspeção de dendrogramas. Isto se deve ao fato de haver dificuldade em se encontrar na literatura um critério objetivo de fácil aplicação para identificar o número ideal de grupos formados. Diante deste problema, o presente trabalho teve por objetivos: 1) Avaliar a aplicabilidade de critério objetivo de se obter o ponto de corte (número ótimo de *clusters*) num dendrograma para a tomada de decisão; 2) trabalhar os conceitos de índices como RMSSTD (root mean square standard deviation) e RS (R-Squared), discutindo a contribuição de cada um destes na obtenção do número ótimo de c*luster*s em acessos de *Capsicum chinense*; 3) aplicação do método, visando a identificar acessos divergentes de *Capsicum chinense* para serem utilizados em programas de melhoramento. Os índices RMSSTD e RS são calculados de acordo com as

variáveis entre e dentro dos grupos formados, caracterizando uma forma objetiva para determinar o número ótimo. Para se obter o ponto de máxima curvatura da trajetória dos índices RMSSTD e RS em função do aumento do número de grupos (X), utilizou-se o Método da Máxima Curvatura Modificado. Foram analisadas, por meio da análise de agrupamentos, algumas características morfológicas de quarenta e nove acessos da espécie *Capsicum chinense* Jacq. do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa. A partir das técnicas propostas agrupou-se os acessos, obtendo um número ótimo de grupos. Os resultados classificam os 49 acessos avaliados em apenas sete grupos de acordo com o gráfico do RMSSTD versus o número de grupos e o gráfico do RS versus o número de grupos.

ABSTRACT

FARIA, Priscila Neves, M.Sc., Universidade Federal de Viçosa, January, 2009. **Evaluation of methods for determining the optimal number of clusters in the study of the genetic divergence among pepper accessions.** Adviser: Paulo Roberto Cecon. Co-Advisers: Fabyano Fonseca e Silva and Cosme Damião Cruz.

Many times, the interpretation of the results in cluster analysis is done subjectively, that is, through inspection on dendograms, since there are no objective criteria to identify the formed clusters. In face of such a problem, the present study aimed to: (1) find out an objective way to achieve the cut-point (optimal number of clusters) in a dendogram in order to help on taking the right decision; (2) work out index concepts such as Root Mean Square Standard Deviation (RMSSTD) and R Squared (RS), explaining the contribution of each one of them in determining the optimal number of cluster; (3) method application, aiming to identify divergent accessions that will be used on improvement programs. An alternative solution for this problem is to use the RMSSTD and RS which are calculated according to the number of variables among and within the clusters formed, characterizing an objective way to determine the optimal number. Another solution is achieved by using

the RS. Some morphological characteristics of the forty nine accessions of the species *Capsicum chinense* Jacq. from the Germplasm Bank of Vegetables of the Federal University of Viçosa (Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa, Minas Gerais – Brazil) were analyzed by means of cluster analysis. The accessions were clustered based on the proposed techniques and an optimal number of clusters was achieved. The 49 accessions analyzed were classified into only seven clusters according to the graph of the RMSSTD versus the number of clusters and the graph of the RS versus the number of clusters.

1. INTRODUÇÃO

A espécie de pimenta *Capsicum chinense* Jacq. apresenta grande importância na agricultura brasileira, pois geralmente é cultivada em pequenas propriedades nas quais se utiliza mão-de-obra familiar. Seus frutos são muito utilizados para confecção de condimentos e especiarias, e apresenta propriedades farmacêuticas, tais como anestésico e antiinflamatório. Porém, há carência de variedades melhoradas com características de cor, aroma e pungência de fruto que atendam as necessidades da indústria, dificultando o desenvolvimento de um mercado organizado e industrial (BOSLAND; VOTAVA, 2000).

O desenvolvimento de cultivares por meio do melhoramento genético possibilitaria a expansão do cultivo da pimenta de forma a atender a demanda industrial. Contudo, o sucesso em programas de melhoramento de plantas depende da diversidade genética do material utilizado, pois quanto maior esta diversidade, maior a probabilidade de se detectar indivíduos superiores, ou seja, com as características comerciais de interesse. O Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa (BGH) possui cerca de cem acessos de *C. chinense* com grande potencial para contribuir em programas de melhoramento, uma vez que grandes quantidades de informações a respeito de

características agronômicas, químicas, morfológicas e moleculares já foram coletadas.

Para avaliar a diversidade genética de forma simultânea em relação a todas estas características, recomenda-se a utilização de medidas de dissimilaridade (CRUZ; CARNEIRO, 2003). Uma forma prática e eficiente de se obter estas medidas é por meio da análise de agrupamentos (Análise de Cluster – AC), a qual tem por finalidade reunir as variáveis em grupos, de forma que exista máxima homogeneidade dentro do grupo e máxima heterogeneidade entre os grupos (JOHNSON; WICHERN, 1992; CRUZ; REGAZZI, 2001). Porém, uma dificuldade desta análise é a falta de critérios objetivos para identificar o número ideal de grupos formados, uma vez que na prática este número é dado simplesmente por uma inspeção gráfica visual.

De acordo com o exposto, o objetivo do presente trabalho é propor critérios objetivos para a determinação do número ótimo de clusters na análise de divergência genética. Será usado como modelo de aplicação as informações de acessos de *Capsicum chinense*. Tais critérios serão fundamentados na obtenção do ponto de máxima curvatura para as trajetórias dos índices RMSSTD (Root Mean Square Standard Deviation) e RS (R-Squared) dada em função do número de clusters.

2. REVISÃO DE LITERATURA

2.1. Diversidade Genética

A definição de diversidade genética foi postulada como "qualquer medida quantitativa ou diferença genética, estando ao nível de seqüência ou nível de freqüência alélica, que é calculada entre indivíduos, populações ou espécies" (BEAUMONT et al., 1998; MOHAMMADI; PRASANNA, 2003).

De acordo com Yorinori e Kiihl (2001), a diversidade genética é a maior garantia da estabilidade de produção, da produtividade e da sobrevivência da humanidade. Porém, a evolução da produtividade agrícola tem sido baseada, principalmente, na uniformidade genética, colocando a atividade agrícola em situação de contínua vulnerabilidade genética e risco de perdas por doenças.

A diversidade genética pode ser analisada em termos de genótipos individuais, linha pura, clones e espécies silvestres. Geralmente o estudo de diversidade entre indivíduos tem por objetivo a identificação de genótipos de interesse específico, a gestão de recursos genéticos e/ou a divisão dos genótipos em grupos heteróticos (SOUZA, 2001).

Há duas maneiras básicas de se inferir sobre a diversidade genética, sendo a primeira de natureza quantitativa e a outra de natureza preditiva

(CRUZ; CARNEIRO, 2003). Os diversos conjuntos de dados têm sido utilizados para estudar a diversidade genética nas culturas, dentre esses conjuntos de dados o mais importante tem sido os dados de pedigree; posteriormente os dados morfológicos; dados bioquímicos; e mais recentemente dados baseados em marcadores de DNA (MOHAMMADI; PRASANNA, 2003). Os diversos conjuntos de dados fornecem variáveis quantitativas, variáveis binárias e variáveis multicategóricas, dentre outras.

Os métodos preditivos de diversidade genética têm sido bastante utilizados, sobretudo pelo fato de que, ao se basearem em diferenças morfológicas, fisiológicas e moleculares dos genótipos, dispensam a obtenção das combinações híbridas entre eles, o que é vantajoso, especialmente quando o número de genitores cujas diversidades se desejam conhecerem é elevado (CARVALHO et al., 2003). Por esses métodos as informações múltiplas de cada cultivar são expressas em medidas de dissimilaridade, que representam a diversidade existente no conjunto de acessos estudados.

As medidas de dissimilaridade comumente utilizadas em variáveis quantitativas são: distância euclidiana; distância euclidiana média; distância euclidiana média padronizada, quadrado da distância euclidiana média; distância generalizada de Mahalanobis (que apresenta certa vantagem visto que levam em consideração as variâncias e covariâncias residuais existentes entre as características mensuradas), entre outras. Em variáveis binárias (dados moleculares) pode-se utilizar: coeficiente de Jacard; coeficiente de Nei e Li; coeficiente de coincidência simples, dentre outros. O procedimento de avaliar a diversidade genética entre acessos a partir de dados moleculares utiliza variáveis binárias, sendo avaliadas a presença e ausência de marcas. Já para as variáveis multicategóricas, isto é, de características morfológicas atribuídas à estrutura de planta, assim como, atributos que conferem qualidade aos produtos comercializados como forma cor e sabor, são comumente determinadas utilizando-se a distância de Cole-Rodgers et al. (1997), onde as características, que normalmente não podem ser ordenadas, são classificadas

em escalas, podendo então ser analisados como características quantitativas discretas (CRUZ; CARNEIRO, 2003).

As estimativas de dissimilaridade atendem aos objetivos do melhorista, por quantificarem e informarem sobre o grau de semelhança ou de diferença entre pares de indivíduos. Entretanto, quando o número de acessos é relativamente grande, torna inviável o reconhecimento de grupos homogêneos pelo exame visual das estimativas de distância. Devido a isso, os acessos semelhantes são reunidos com o uso de técnicas de agrupamento ou de projeções de distâncias em gráficos bidimensionais, em que cada coordenada é obtida a partir da medida de dissimilaridade escolhida. A união se dá pela classificação dos acessos em vários grupos, de forma que exista homogeneidade dentro e heterogeneidade entre os grupos. Ou seja, o grupo original é dividido em vários grupos seguindo o critério de similaridade ou de dissimilaridade (CRUZ; CARNEIRO, 2003).

2.2. Análise de Agrupamento

A técnica multivariada de análise de agrupamento (Análise de *Clusters* - AC) é uma maneira de se obter grupos homogêneos, por um esquema que possibilite reunir os acessos em questão em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles (JONHSON; WICHERN, 1992; CRUZ; REGAZZI, 2001).

De acordo com Maxwell (1977), o primeiro passo da análise de agrupamento é a conversão da matriz $n \times p$ de dados, em outra matriz $n \times n$ de medidas de similaridade ou dissimilaridade, mensurada em relação aos pares de n unidades amostrais, em função de um conjunto de "p" características (variáveis). De posse das estimativas de distância entre cada par de elementos estudados, os dados são apresentados em uma matriz simétrica $n \times n$, e a partir desta, a visualização e interpretação das distâncias pode ser facilitada pela utilização de um método de agrupamento e/ou dispersão gráfica.

Vários trabalhos têm utilizado a análise de agrupamentos no estudo da diversidade genética. Carvalho et al. (2003) analisaram a diversidade genética entre acessos do banco ativo de germoplasma de algodão, detectando grande diversidade genética entre os acessos, que foram reunidos em dez grupos distintos. Sudré et al. (2005) analisaram a divergência genética entre acessos de pimenta e pimentão utilizando técnicas multivariadas, observando-se concordância entre todas as técnicas utilizadas e a separação dos acessos em oito grupos distintos, indicando a existência de variabilidade genética entre os acessos.

Lannes (2005) avaliou a diversidade genética em *Capsicum chinense* mediante estudo de características químicas, morfológicas e moleculares. Este autor utilizou a análise de agrupamentos considerando a distância de Mahalanobis e o método de otimização de Tocher, e os resultados apontaram para a formação de dez grupos.

Karasawa et al. (2005) estudaram a divergência genética entre acessos de tomateiro, e relataram que a utilização do procedimento subjetivo, baseado no exame visual do dendrograma, pode gerar alguma dificuldade na tomada de decisão quanto ao número de grupos gerados, uma vez que qualquer inferência rígida sobre este número pode não ser produtiva.

2.2.1. Medidas de Dissimilaridade para variáveis contínuas

Após o pesquisador definir as características as quais se tem interesse em avaliar, espera-se que a distância entre os indivíduos não seja alterada com a adoção de características com unidades de medidas distintas, num mesmo conjunto de dados. É importante que as variáveis admitidas apresentem poder discriminatório semelhante, não baseado na amplitude de seus valores. Caso uma determinada unidade de medição apresente uma maior amplitude em seus valores, em comparação às demais, ela certamente terá um maior peso na análise.

Assim, se todos os dados estiverem num mesmo padrão (unidade) de medida, pode-se neste caso, utilizar os dados originais. Entretanto, a não ocorrência deste mesmo padrão de medida entre as variáveis sugere uma padronização dos dados.

A padronização dos dados mais difundida segue a seguinte estratégia:

$$Z_j = \frac{X_{ij}}{\hat{\sigma}_j}$$

em que

X_{ij} é a média do i-ésimo indivíduo para a j-ésima característica e $\hat{\sigma}_j$ é o desvio-padrão associado à j-ésima característica.

Existem duas razões para a padronização dos dados. A primeira visa evitar que as unidades escolhidas para mensurar as características afetem arbitrariamente a similaridade entre os indivíduos. A segunda é que a padronização faz com que as características contribuam igualmente na avaliação da similaridade entre indivíduos.

Um grande número de medidas de similaridade ou de dissimilaridade tem sido proposto e utilizado em análise de agrupamento, sendo a escolha entre elas baseada na preferência e, ou, na conveniência do pesquisador (BUSSAB et al., 1990).

O termo dissimilaridade surgiu em função da relação da distância entre dois pontos P e Q, definida como d(P,Q), pois, à medida que ela cresce, diz-se que a divergência entre os pontos (unidades amostrais) P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares.

Os valores de distâncias são geralmente obtidos a partir de informações de “n” unidades amostrais, mensurados em relação a “p” caracteres (variáveis).

É necessário especificar um coeficiente de semelhança que indique a proximidade entre os indivíduos sendo importante considerar, em todos os casos semelhantes a este, a natureza da variável (discreta, contínua, binária) e a escala de medida (nominal, ordinal, real ou razão).

Conforme descrito por Johnson e Wichern (1992) e Mardia et al. (1997), cita-se como medida de dissimilaridade a Distância Euclidiana**,** que é insatisfatória para muitas situações estatísticas. Isso ocorre devido à contribuição de cada coordenada ter o mesmo peso para o cálculo da distância. Arunachalam (1981) preconiza que, em estudos sobre distanciamento genético, somente é aconselhável a quantificação da dissimilaridade pela distância Euclidiana, quando forem avaliadas várias características cujos graus de correlação residual não sejam significativos. Entretanto, como em estudos de melhoramento é praticamente impossível avaliar um conjunto de características não-relacionadas, o uso da distância Euclidiana tem sido indiscriminado. Porém, tem-se mostrado de grande utilidade mesmo nas situações em que a independência entre as características mensuradas não é constatada. Essa medida foi utilizada neste trabalho por ser uma das mais usadas na prática e pela facilidade de ser encontrada nos mais diversos programas computacionais.

Outra distância muito utilizada é a Distância de Mahalanobis. Esta distância considera as diferenças de variação e a presença de correlação, ou seja, é uma distância que depende das variâncias e das covariâncias amostrais. Uma outra medida de distância é a Métrica de Minkowski, a qual depende de funções modulares. Existem ainda a Distância Euclidiana Média e a Distância Euclidiana Padronizada. Outros tipos de definições de distâncias podem ser encontrados na literatura (KHATTREE; NAIK, 2000; BUSSAB et al., 1990).

2.2.2. Medidas de dissimilaridade para variáveis discretas

Muitas vezes os objetos não podem ser mensurados em variáveis quantitativas e então, essas variáveis podem ser transformadas em dicotômicas (binárias). As variáveis qualitativas podem ser transformadas em variáveis binárias tomando-se como valor 1 a presença de uma determinada realização e o valor 0 para as demais. Desta forma, existem os coeficientes de

dissimilaridade (ou similaridade), que são utilizados para variáveis qualitativas.

O coeficiente de similaridade indica a força de relação entre os indivíduos ou variáveis, fixando um valor comum aos mesmos (EVERITT, 1993). Existem técnicas de agrupamentos que vem sendo propostas. No entanto, o mais comumente usado, segundo Landim (2001), é o agrupamento pareado não ponderado baseado na média aritmética (unweighted pair-group method using arithmetic averages "UPGMA"), que realiza o cálculo dos valores médios das variáveis e atribui sempre o mesmo peso aos dois elementos que estão sendo integrados.

Os coeficientes de similaridade mais usuais, obtidos num espaço multidimensional, podem ser subdivididos entre os que medem a distância ou a separação angular entre pares de pontos; entre os que medem a correlação entre pares de valores; e entre os que medem a associação entre pares de caracteres qualitativos. Existem diversas publicações que discutem esses diversos tipos de medidas como, por exemplo, Sneath e Sokal (1973) e Everitt (1993).

2.2.3. Métodos de agrupamento

Há inúmeros métodos de agrupamento, que se distinguem pelo tipo de resultado a ser fornecido e pelas diferentes formas de definir a proximidade entre um indivíduo e um grupo já formado ou entre dois grupos quaisquer.

O teste aglomerativo de Scott Knott (1974) visa a separação de médias de tratamentos em grupos distintos, através da minimização da variação dentro e maximização da variação entre grupos. Os resultados são facilmente interpretados, devido à ausência de ambigüidade.

Dentre os métodos aglomerativos, os mais utilizados são os de otimização e os hierárquicos. Nos métodos de otimização os grupos são formados pela adequação de algum critério de agrupamento, ou seja, o objetivo é alcançar uma partição dos indivíduos que otimize (maximize ou

minimize) alguma medida pré-definida. Um dos métodos mais comumente utilizado na área de melhoramento genético é o proposto por Tocher, citado por Rao (1952). Nos métodos hierárquicos, os indivíduos são agrupados por um processo que se repete em vários níveis até que seja estabelecido o dendrograma ou o diagrama de árvore. Neste caso, não há preocupação com o número ótimo de grupos, uma vez que o interesse maior está na "árvore" e nas ramificações que são obtidas.

As técnicas hierárquicas são as mais amplamente difundidas (SIEGMUND et al., 2004) e envolvem basicamente duas etapas. A primeira se refere à estimação de uma medida de similaridade ou dissimilaridade entre os indivíduos e a segunda, à adoção de uma técnica de formação de grupos (SANTANA; MALINOVSKI, 2002).

Segundo Cruz e Regazzi (2001), existem várias formas de representar esta estrutura de agrupamento, tais como: o método do vizinho mais próximo, o método do vizinho mais distante, método UPGMA (agrupamento pareado não ponderado baseado na média aritmética), método de Ward, dentre outros.

O método UPGMA é o mais utilizado em diversidade quando se trabalha com populações silvestres e tendo vantagem sobre os demais métodos por considerar médias aritméticas das medidas de dissimilaridade, o que evita caracterizar a dissimilaridade por valores extremos entre os indivíduos considerados (CRUZ; CARNEIRO, 2003). É um método de agrupamento seqüencial, aglomerativo, hierárquico, sem superposição e com base na média aritmética. Neste método, a distância entre dois agrupamentos é a distância média entre todos os pares de observações, um em cada agrupamento. Esse método foi utilizado por ser um dos mais usados na prática e pela facilidade de ser encontrado nos mais diversos programas computacionais.

2.2.4. Dendrograma

Os agrupamentos são feitos utilizando todas as variáveis disponíveis e representados de maneira bidimensional através de um dendrograma

(diagrama bidimensional em forma de árvore). Nele estão dispostos linhas ligadas segundo os níveis de similaridade (ou dissimilaridade), que agrupará indivíduos ou grupos de indivíduos (EVERITT, 1993; LANDIM, 2001).

O dendrograma ilustra as fusões ou partições efetuadas em cada nível sucessivo do processo de agrupamento, no qual um eixo representa os indivíduos e o outro eixo representa as distâncias obtidas após a utilização de uma metodologia de agrupamento. Os ramos da árvore fornecem a ordem das (n-1) ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

De forma geral, os dendrogramas apresentam estruturas de agrupamentos de objetos homogêneos. Entretanto, a falta de critérios objetivos para se determinar o ponto de corte no dendrograma (número ótimo de grupos) ainda é um problema em estudos que utilizam a análise de agrupamentos. Um método considerado como "objetivo", dentre os poucos existentes, é o Método de Mojema (1977). Este Método é um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendrograma.

No presente trabalho, propõe-se ainda outro critério, de fácil entendimento, para determinação do número ótimo de grupos, baseado nas trajetórias das curvas dos índices RMSSTD e RS, utilizando o Método da Máxima Curvatura Modificado. Este último é utilizado para determinar o ponto de curvatura máxima das referidas curvas, ponto este que determina o número ótimo buscado.

2.3. Índice RMSSTD e RS

O índice RMSSTD (Root Mean Square Standard Deviation), cuja tradução pode ser raiz-quadrada do desvio padrão médio, é usado para calcular a homogeneidade dos agrupamentos (SHARMA, 1996). Em outras palavras, quanto mais compactos foram os grupos formados, situação esta verificada na presença de um grande número de grupos, menores os valores para esta estatística. Assim, é possível visualizar um gráfico (Figura 1a) que mostra o

decréscimo do RMSSTD em função do aumento do número de clusters, todavia, esta trajetória não é linear, e o seu ponto de máxima curvatura indica um limiar entre uma fase de decréscimo e uma fase de estabilização. Após este ponto, denominado de ótimo, mesmo aumentando o número de clusters não se verifica grandes declínios nos valores do RMSSTD.

Com relação ao índice R-square (RS), ou coeficiente de determinação, este é usado para calcular a dissimilaridade entre agrupamentos. Um alto valor de RS indica dissimilaridade mais alta entre grupos (SHARMA, 1996), e tal situação é representada na presença de um alto número de grupos. Graficamente, o aumento no número de cluster proporciona um aumento nos valores do RS (Figura 1b), e esta trajetória também é não-linear, o que realça a importância de se calcular um ponto de máxima curvatura.

Na literatura estatística, pontos de máxima curvatura geralmente são calculados em estudos de determinação de tamanho ótimo de parcelas experimentais, portanto os métodos empregados nesta ocasião podem ser usados para estimar o número ótimo de cluster em se tratando das trajetórias mostradas na Figura 1.

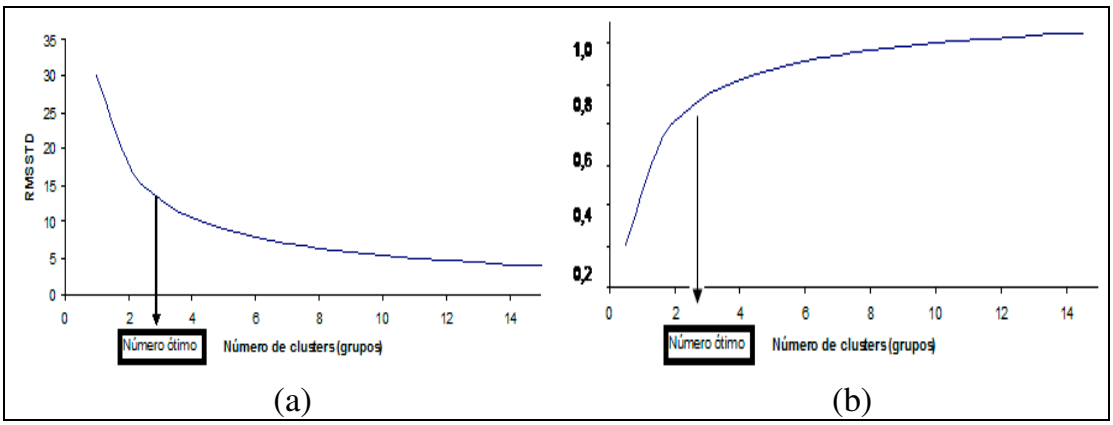


Figura 1 – Trajetória dos índices RMSSTD (a) e RS (b) em função do aumento do número de clusters (grupos).

2.4. Método da Máxima Curvatura Modificado

A expressão de Smith (1938) para o Método da Máxima Curvatura é apresentada na forma exponencial e relaciona o coeficiente de variação, CV, e o tamanho da amostra, conforme a equação:

$$CV = \frac{a}{X^{b}}$$

em que

a e b são as constantes apropriadas;

CV é o coeficiente de variação por unidade básica;

X é o número de unidades básicas.

O Método da Máxima Curvatura Modificado foi proposto por Lessman e Atkins (1963), para que o ponto de máxima curvatura não dependa da escala das coordenadas, mas do tamanho ótimo da parcela.

Lessman e Atkins (1963), calcularam os CV e a partir deles estimaram as constantes apropriadas, a e b. Derivando a equação proposta em ordem a X, obtiveram as tangentes aos vários pontos da curva. As duas tangentes sucessivas com o maior ângulo θ de interseção definem a região de curvatura máxima, na qual a taxa de mudança em CV é a maior em relação aos aumentos em X. A derivada de θ em ordem a X permite obter o valor de X ou o tamanho ótimo da parcela. Lessman e Atkins (1963) apresentam a seguinte expressão para a determinação do ponto crítico ou tamanho ótimo da parcela:

$$\text{Xcrítico} = 2b + 2\sqrt{a^{2}b^{3}/(b+1)}$$

em que

X, a e b são como definidos anteriormente.

Aplicando estes conceitos a ensaios de uniformidade de sorgo, Lessman e Atkins (1963) constataram que o CV diminui progressivamente com o aumento do tamanho da parcela, mas tende a estabilizar-se com o aumento das parcelas. Os resultados obtidos com o método da Curvatura Máxima Modificado foram muito semelhantes aos obtidos com o método de Smith

(1938). A aplicação à produção de sorgo e, talvez à de grãos, parece ter um ajuste melhor usando a equação:

$$y = a/(1 + \log X)^{b}$$

em que

y é o CV e X, a e b são idênticos aos anteriormente definidos.

O tamanho ótimo da parcela estimado das duas maneiras é praticamente o mesmo (LESSMAN; ATKINS, 1963).

O valor de X obtido pela expressão de Lessman e Atkins (1963) apresenta um viés em direção a valores menores de X, segundo Meier e Lessman (1971), os quais propõem a seguinte expressão para estimar o ponto de curvatura máxima:

$$Xc = \left[a^{2} b^{2} (2b + 1)/(b + 2)\right]^{1/(2b+2)}$$

em que

X_c é o ponto de curvatura máxima e as outras letras são como acima definidas.

O tamanho da parcela obtido com esta modificação é 20% menor que o resultante do método de Smith (1938). Como diferenças de até 50%, para mais ou para menos, no tamanho das parcelas pouco afetam a respectiva eficiência, esta modificação da metodologia terá pequena influência nos resultados finais (MEIER; LESSMAN, 1971).

VIANA (1999) ressalta que o método da máxima curvatura modificado fornece resultados mais precisos, pois estabelece uma equação de regressão para explicar a relação entre os coeficientes de variação e os respectivos tamanhos de parcelas.

Esta metodologia foi adaptada ao estudo deste trabalho e utilizamos Xc como sendo o ponto de máxima curvatura da trajetória dos índices RMSSTD e RS em função do aumento do número de grupos (X).

3. MATERIAL E MÉTODOS

3.1. Descrição dos dados

Foram avaliados quarenta e nove acessos de *C. chinense* (Tabela 1), pertencentes ao Banco de Germoplasma de Hortaliças (BGH) da Universidade Federal de Viçosa (UFV), quanto a características quantitativas que conferem qualidade aos frutos da planta.

O experimento foi conduzido no delineamento de blocos ao acaso, com três repetições, utilizando espaçamento de 1 m x 1 m entre plantas e linhas, sendo cada linha constituída por três plantas de cada acesso. Foram analisadas as seguintes características:

comp (mm): Comprimento;

larg (mm): Largura;

% MS: Porcentagem de matéria seca;

mst (g): massa da matéria total do fruto maduro seco ;

peso (g): massa da matéria total do fruto maduro fresco;

capT (mg.g MS^{-1}): Capsaicina total;

tss (^{o} Brix): Teor de Sólidos solúveis;

vit C (mg/100 g de fruto fresco): Vitamina C;

cor ext (unid. ASTA de cor): Cor extraível;

Esp. (cm): espessura do pericarpo.

A fim de facilitar a interpretação, serão identificados os 49 acessos numerando-os de 1 a 49. A numeração (N°) dos acessos a que serão referidos daqui em diante estão na Tabela 1.

Tabela 1 – Relação dos quarenta e nove acessos de *C. chinense* selecionados do Banco de Germoplasma de Hortaliças da UFV e sua respectiva numeração (N°)

Acesso	N°	Acesso	N°	Acesso	N°	Acesso	N°
BGH1694-05	**1**	BGH1747-26	**13**	BGH4733-54	**25**	BGH6228-82	**37**
BGH1694-06	**2**	BGH1747-27	**14**	BGH4733-55	**26**	BGH6233-83	**38**
BGH1694-07	**3**	BGH4199-30	**15**	BGH4733-56	**27**	BGH6233-84	**39**
BGH1714-09	**4**	BGH4201-32	**16**	BGH4744-57	**28**	BGH6233-85	**40**
BGH1714-11	**5**	BGH4213-34	**17**	BGH4750-59	**29**	BGH6239-86	**41**
BGH1716-14	**6**	BGH4223-39	**18**	BGH4756-67	**30**	BGH6344-87	**42**
BGH1716-16	**7**	BGH4285-40	**19**	BGH4756-70	**31**	BGH6369-90	**43**
BGH1716-17	**8**	BGH4289-44	**20**	BGH4756-71	**32**	BGH6371-93	**44**
BGH1716-18	**9**	BGH4289-45	**21**	BGH5012-72	**33**	BGH6371-94	**45**
BGH1716-19	**10**	BGH4355-46	**22**	BGH5012-76	**34**	BGH6371-95	**46**
BGH1723-22	**11**	BGH4725-51	**23**	BGH6009-78	**35**	BGH6378-98	**47**
BGH1724-23	**12**	BGH4731-53	**24**	BGH6228-79	**36**	BGH6387-100	**48**
						BGH7295-101	**49**

Em relação aos dados utilizados na análise de agrupamentos, eles são uma coleção de informações sobre os 49 indivíduos ou unidades. Há dois formatos comuns nos quais os dados podem ser apresentados, ambos envolvem a noção de uma matriz: uma matriz de dados ou uma matriz de dissimilaridade. A matriz de dados utilizada na análise de agrupamentos deste trabalho contém as médias de 10 variáveis (características) para cada um dos 49 acessos.

Por convenção, o número de indivíduos é igual ao número de linhas (49) considerando que o número de variáveis é igual ao número de colunas (10). Utilizou-se neste trabalho as técnicas que agrupam indivíduos para medidas quantitativas.

3.2. Análise de variância individual e estimação de parâmetros

Foi realizada análise de variância dos 49 tratamentos das características avaliadas com base na média das parcelas, visando avaliar a existência de variabilidade genética significativa entre os tratamentos, onde será utilizado o modelo estatístico abaixo

$$Y_{ij} = \mu + G_i + \beta_j + \varepsilon_{ij}$$

em que

Y_{ij} = valor fenotípico da ij-ésima observação referente ao i-ésimo tratamento no j-ésimo bloco;

μ = média geral do caráter;

G_i = efeito do i-ésimo tratamento (i = 1, 2, 3, ... , g; g = 49);

β_j = efeito do j-ésimo bloco (j = 1, 2, ... , r; r = 3); e

ε_{ij} = efeito do erro experimental, sendo ε_{ij} NID (0, σ^2)

Como os resultados obtidos serão válidos apenas para os materiais genéticos em questão, considerar-se-á o modelo misto. Assim, a hipótese testada pela estatística F será Ho: $G_i = 0$, para todo i.

O esquema da análise de variância com as esperanças dos quadrados médios [E(QM)], considerando o efeito de tratamento como fixo, segundo Steel e Torrie (1980), está apresentado no Quadro 1.

Quadro 1 – Esquema de análise de variância e esperanças de quadrados médios de um modelo em blocos casualizados, com efeitos de tratamentos fixos

FV	GL	QM	E(QM)	F
Blocos	r – 1	QMB	$\sigma^2 + g\Phi_b$	
Tratamentos(G)	g – 1	QMG	$\sigma^2 + r\Phi_g$	QMG/QMR
Resíduo	(r –1)(g – 1)	QMR	σ^2	

em que

σ^2 = componente de variância devido ao erro experimental;

Φ_b = componente de variância devido ao bloco; e

Φ_g = componente quadrático associado aos tratamentos.

sendo $\Phi_g = \frac{\sum_{i=1}^{g} G_i^2}{g-1}$.

3.3. Agrupamento dos acessos

A diversidade genética entre os acessos de *C. chinense* foi avaliada pelo método UPGMA (agrupamento pareado não ponderado baseado na média aritmética) e as medidas de dissimilaridade entre acessos considerando variáveis quantitativas foram determinadas pela distância Euclidiana, mediante o emprego do software estatístico SAS® (SAS, 1999).

Considerando X_{ij} a média no i-ésimo indivíduo (clone, cultivar, linhagem etc.) para a j-ésima característica, define-se a distância Euclidiana entre o par de indivíduos i e i' por meio da expressão:

$$d_{ii'} = \sqrt{\sum_j \left(X_{ij} - X_{i'j}\right)^2}$$

3.3.1. Método UPGMA

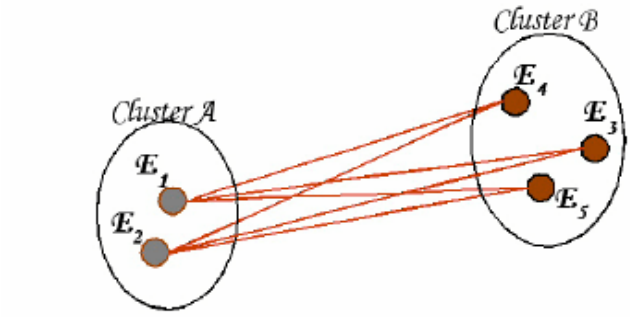
Neste método o dendrograma é estabelecido pelos indivíduos com maior similaridade, sendo que a distância entre um indivíduo k e um grupo formado pelos indivíduos i e j é dada por:

$$d_{(ij)k} = média\{d_{ik}; d_{jk}\} = \frac{d_{ik} + d_{jk}}{2},$$

onde $d_{(ij)k}$ é a média do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k). A distância entre os dois agrupamentos é definida por:

$$d_{(ij)(kl)} = média\{d_{ik}; d_{il}; d_{jk}; d_{jl}\} = \frac{d_{ik} + d_{il} + d_{jk} + d_{jl}}{4},$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é a média do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e k), (i e l) , (j e k) e (j e l). Na Figura 2 (METZ; MONARD, 2006), um esquema do que ocorre:



$$d_{AB} = (d(E_1, E_3) + d(E_1, E_4) + d(E_1, E_5) + d(E_2, E_3) + d(E_2, E_4) + d(E_2, E_5)) / 6$$

Figura 2 – Average Link (UPGMA): média das distâncias entre dois clusters.

Este método foi utilizado por ser um dos mais difundidos e relevantes, face sua importância em estudos relacionados ao melhoramento genético, calculado com base na distância entre os pares de unidades amostrais.

3.3.2. Método de otimização de Tocher

O método de otimização de Tocher se baseia na identificação do par mais similar dentro da matriz de dissimilaridade, isto é, aquele com menor estimativa de distância. Esses indivíduos formarão o primeiro grupo e a partir desse, é avaliada a possibilidade de inclusão de novos indivíduos no grupo, adotando o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo (CRUZ; CARNEIRO, 2003).

A inclusão, ou não, do individuo k no grupo é, então, feita considerando:

- Se $\frac{d(grupo)k}{n} \leq \theta$, inclui-se o individuo K no grupo;

- Se $\frac{d(grupo)k}{n} > \theta$, não se inclui o individuo K no grupo;

Sendo: n é o número de indivíduos que constitui o grupo original,

θ é o valor máximo da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo.

Nesse caso, a distância entre o individuo k e o grupo formado pelos indivíduos ij é dada por:

$$d_{(ij)k} = d_{ik} + d_{jk}$$

Foi utilizado o aplicativo computacional GENES (CRUZ, 2001) na obtenção do número de grupos pelo Método de Tocher.

3.4. Número ótimo de clusters

Foi utilizado o método da Método da Máxima Curvatura Modificado (LESSMAN; ATKINS, 1963), que é baseado na modificação do método de Smith (1938).

No presente estudo, tendo em vista os índices RMSSTD e RS, o modelo usado foi dado por: $RMSSTD = \frac{a}{X^{b}}$ e $RS = \frac{a'}{X^{b'}}$, em que X representa o número de clusters, *a, b* e *a', b'* são constantes apropriadas.

A partir da função de curvatura, determinou-se o valor da abscissa onde ocorre o ponto de máxima curvatura, conforme apresentado por Meier e Lessman (1971), por meio de:

$$X_{MC} = \left[\frac{a^2 b^2 (2b+1)}{(b+2)} \right]^{\frac{1}{2b+2}}$$

em que *a* e *b* são constantes apropriadas.

De forma geral, algoritmos de agrupamentos hierárquicos normalmente usam o índice RMSSTD, mas ele pode ser usado para avaliar os resultados de qualquer algoritmo de agrupamento. O RMSSTD (raiz-quadrada do desvio padrão médio) é usado para determinar o número de agrupamentos inerente a um conjunto de dados, medindo a homogeneidade dos agrupamentos resultantes. O valor do índice deve ser tão baixo quanto possível para um agrupamento, isto é, quanto menor o RMSSTD, mais homogêneo ou compacto é o agrupamento formado a um determinado passo. Um valor grande de RMSSTD sugere que o agrupamento obtido a um determinado passo não é homogêneo. Assim, mais baixo valor do RMSSTD significa agrupando melhor e seu valor é expresso pela equação:

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} (n_{ij} - 1)}}$$

em que: $i = 1,...,nc$ e $j = 1,...,d$.

nc é o número de grupos, d é o número de variáveis, $\bar{x}_j$ é o valor esperado na *j*-ésima variável, n_{ij} número de elementos no i-ésimo grupo na j-ésima variável e n_j é o número de elementos na j-ésima variável em todo o conjunto de dados.

O R-square (RS) mede a heterogeneidade da solução do agrupamento formada a um determinado passo (mede a diferença entre os agrupamentos).

Um valor grande representa que os agrupamentos obtidos a um determinado passo são bastante diferentes (heterogêneos) um do outro, e um valor pequeno significa que os agrupamentos formados a um determinado passo não são muito diferentes um do outro. Assim, o índice RS descrito acima é a medida de dissimilaridade de agrupamentos. Temos que $0 \leq RS \leq 1$, e se seu valor for 0, não há nenhuma diferença entre os agrupamentos; se o seu valor é 1 então há uma diferença entre eles.

RS pode ser definido da seguinte forma:

$$RS = \frac{\left[\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_j} \left(x_i - \bar{x}_j\right)^2\right] - \left[\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} \left(x_k - \bar{x}_j\right)^2\right]}{\sum_{j=1\ldots d} \sum_{k=1}^{n_j} \left(x_k - \bar{x}_j\right)^2}$$

em que: $i = 1,\ldots,nc$ e $j = 1,\ldots,d$.

nc é o número de grupos, d é o número de variáveis, $\bar{x}_j$ é o valor esperado na *j*-ésima variável, n_{ij} número de elementos no i-ésimo grupo na j-ésima variável e n_j é o número de elementos na j-ésima variável em todo o conjunto de dados.

Para empregar a metodologia proposta, inicialmente realizou-se uma análise multivariada utilizando o software SAS® (SAS, 1999) utilizando o PROC GLM com a opção MANOVA. Após esta análise, foram também extraídas as médias (LSMEANS) de cada variável em cada grupo formado. Posteriormente, os dados foram submetidos à análise de agrupamento no procedimento PROC CLUSTER (SAS, 1999) pelo método UPGMA. Nesta última análise, obtiveram-se os valores do RMSSTD e RS em relação ao número de grupos, gerando gráficos que possibilitaram a identificação do número ótimo de grupos em relação à máxima curvatura, a qual foi observada geometricamente (LARSON et al., 1998).

Foi utilizado o PROC NLIN (SAS, 1999) a fim de obter estimativas para os parâmetros *a* e *b* do modelo proposto pelo método da máxima

curvatura modificado. Para tanto, utilizou-se o método dos Quadrados Mínimos para modelos de regressão não linear via algoritmo de Gauss-Newton.

Como complemento dessa análise, utilizou-se o PROC TREE (SAS, 1999) para visualizar o dendrograma e para verificar quais indivíduos são pertencentes a cada grupo obtido pela discriminação estatística.

Os procedimentos citados se encontram relacionados nos Apêndices 1 e 2.

3.4.1. Método de Mojema (1977)

Mojema (1977) sugeriu um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendrograma. A proposta é selecionar o número de grupos no estágio j que, primeiramente, satisfizer a seguinte inequação: $\alpha_j > \theta_k$

em que

α_j é o valor de distâncias dos níveis de fusão correspondentes aos estágios j (j=1, 2, ..., n), θ_k é o valor referencial de corte, dado por: $\theta_k = \overline{\alpha} + k\hat{\sigma}_\alpha$

Sendo:

$\overline{\alpha}$ e $\hat{\sigma}_\alpha$ são, respectivamente, a média e o desvio-padrão não-viesado dos valores de α.

k é uma constante. A adoção de valores de *k* em torno de 2,75 e 3,50 é sugerida por Mojema (1977). No entanto, Milligan e Cooper (1985) sugeriram o valor de *k* =1,25 como regra de parada na definição do número de grupos.

Assim, tem-se que:

$$\overline{\alpha} = \frac{1}{g-1}\sum_{j=1}^{g-1}\alpha_j \quad \text{e} \quad \hat{\sigma}_\alpha = \sqrt{\frac{\left(\sum_{j=1}^{g-1}\alpha_j^2 - \frac{1}{g-1}\left(\sum_{j=1}^{g-1}\alpha_j\right)^2\right)}{g-2}}$$

4. RESULTADOS E DISCUSSÕES

Pelo teste F ao nível de 1% de probabilidade (Tabela 2) detectou-se diferença significativa entre os acessos para as características avaliadas, indicando a presença de variabilidade genética.

Nas características comprimento, largura, capsaicina total, teor de sólidos solúveis, vitamina C e espessura do pericarpo o coeficiente de variação está entre 5,34% e 18,81%, indicando boa precisão experimental. Quanto à matéria fresca, o coeficiente de variação foi de 29,98%. A massa do fruto seco e a cor extraível apresentaram coeficiente de variação de 33,60% e 36,82% respectivamente (Tabela 2).

As médias das características avaliadas encontram-se nas Tabelas 3 e 4, agrupadas pelo teste de Scott-knott a 5% de probabilidade.

O comprimento dos frutos variou de 14,00 mm a 76,22 mm agrupando os acessos em cinco grupos com 36,73% reunidos no grupo quatro com variação dentro do grupo de 32,78g a 42,31g. Os acessos 44 (BGH 6371-93), 25 (BGH 4733-54), 18 (BGH 4223-39) e 85 (BGH 4285-40) formavam o grupo com maior comprimento de fruto sendo o acesso 44 aquele que apresentava o maior comprimento entre os acessos avaliados.

A largura dos frutos variou de 8,68 mm a 42,89 mm, com a formação de seis grupos. 87% dos acessos possuíam largura inferior a 30 mm e largura. O acesso 35 (BGH 6009-78) possuía maior largura entre os acessos estudados com mais de 42 mm de largura, não formando grupo com os demais acessos.

A porcentagem de matéria seca (%MS) dos acessos variou entre 7,04% a 20,92%. (Tabela 3), formando apenas dois grupos de acordo com teste de Scott-knott a 5% de probabilidade. 30% dos acessos estudados apresentaram porcentagem de matéria seca acima de 14,5%, característica essa de grande importância no rendimento final da indústria de processamento de pimenta em pó. O acesso 27 (BGH 4733-56) possuía a maior porcentagem de matéria seca entre os acessos avaliados.

A espessura de pericarpo (Esp) variou de 1,0 mm a 3,5 mm (Tabela 3). Desta forma, 40% dos acessos estudados apresentavam espessura de pericarpo com mais de 2,0 mm. De acordo com o teste de Scott-Knott os acessos foram agrupados em três grupos sendo o maior, com cerca de 60% dos acessos com espessura menor que 2,0 mm. O acesso 35 (BGH6009-78) possuía o pericarpo mais espesso, entretanto esse acesso agrupou juntamente com o acesso 18 (BGH 4223-39), o acesso 40 (BGH 6233-85), o acesso 47 (BGH 6378-98), o acesso 43 (BGH 6369-90) e com o acesso 3 (BGH 1694-07) com cerca de 3,0 mm de espessura do pericarpo.

O teor de sólidos solúveis variou de 5,37 º Brix a 12,90º Brix (Tabela 4), e pelo teste de Scott-knott os acessos foram reunidos em três grupos, onde cerca de 70% deles possuíam valores menores que 8,5 º Brix. O acesso 27 (BGH 4733-56) possuía o maior teor de sólidos solúveis agrupando juntamente com os acessos 20 (BGH 4289-44) e o acesso 21 (BGH 4289-45) com aproximadamente 13 º Brix.

Os valores quantitativos para intensidades de cor variaram foi 30,35 a 595,84 unidades ASTA de cor (Tabela 4). Pelo teste de Scott-knott os acessos foram agrupados em cinco grupos, sendo esse último o maior com cerca de 50% dos acessos presentes nele. O acesso 27 (BGH 4733-56) apresentou o maior valor de intensidade, sendo que esse acesso terá menor perda de coloração quando da desidratação do fruto. De acordo com o teste de Scott-knott esse acesso não agrupou com os demais acessos.

Baseado no teste de Scott knott a 5% de probabilidade, o teor de capsaicina total foi o que apresentou a maior diversidade entre os acessos, separando eles em 19 grupos. De acordo com a escala Scoville de calor (HSU), o valor máximo de capsaicina total foi de 205000 unidades Scoville de calor do acesso 8 (BGH 1716-17).

Tabela 2 – Resumo da análise da variância de nove características de qualidade de fruto de *C. chinense*, valores do coeficiente de variação e média de cada característica

Fontes de Variação	GL	Quadrado médio									
		Comp +	larg	%MS	mst	peso (mft)	capTot	tss	vit c	cor ext	Esp
Bloco	2	50,77	6,61	2,04	0,012	0, 978	0, 1878	0,3415	20,20	4960,78	0,0002
Acessos	48	761,47**	187,58**	32,23**	0,300**	46,11**	39,20**	5,40**	1315,01**	33198,14**	0,0076**
Resíduo	96	59,90	7,89	7,13	0,047	3,10	0,0681	1,19	286,31	2769,69	0,0013
CV%		18,81	13,45	20,44	33,60	29,98	5,34	13,43	17,32	36,82	17,38
Média		41,14	20,87	13,06	0,649	5,87	4,88	8,11	97,65	142,91	0,2076

[+] comp (mm): Comprimento; larg (mm): Largura; % MS: Porcentagem de matéria seca; mst (g): massa da matéria total do fruto maduro seco ; peso (g): massa da matéria total do fruto maduro fresco ; capTot (mg.g MS^{-1}): Capsaicina total; tss (^{o} Brix): Teor de Sólidos solúveis; vit C (mg/100 g de fruto fresco): Vitamina C; cor ext (unid. ASTA de cor): Cor extraível; Esp. (cm): espessura do pericarpo.

** Significativo a 1% de probabilidade, pelo teste F.

Tabela 3 – Médias de sete características avaliadas em quarenta e nove acessos de *C.chinense*

Acesso	Comp (mm)	Larg (mm)	Cap Tot (mg.g ms^{-1})	%MS	Massa seca total (g)	Massa fresca total(g)	Espessura (cm)
1	27,7800 e*	22,1733 d	4,5328 l	13,9054 b	0,4980 c	3,5553 e	0,1822 c
2	23,4333 e	21,2733 d	4,6978 l	12,9158 b	0,4107 d	3,1780 e	0,1744 c
3	33,4000 d	28,5467 c	1,8654 q	10,2092 b	0,9840 b	9,6527 c	0,2800 a
4	41,8200 d	21,9867 d	2,1363 p	16,7273 a	0,8600 b	5,9180 d	0,2400 b
5	40,3200 d	25,4600 c	2,3621 p	10,7908 b	0,6200 c	5,5847 d	0,2211 b
6	36,8067 d	19,0867 d	0.0003 s	11,6846 b	0,4953 c	4,2787 d	0,2022 c
7	37,7400 d	20,7667 d	0.0003 s	12,4502 b	0,7220 c	6,0600 d	0,2533 b
8	52,2933 c	13,7267 e	13,6273 a	12,5152 b	0,5233 c	4,2793 d	0,2033 c
9	54,6867 c	17,5467 d	6,7702 i	11,8357 b	0,6593 c	5,8020 d	0,1922 c
10	62,8267 b	14,1667 e	1,8494 q	11,6276 b	0,5740 c	5,2927 d	0,1622 c
11	29,5600 e	22,3200 d	4,9081 k	10,4515 b	0,5027 c	4,8107 d	0,1967 c
12	21,7067 e	12,8667 e	8,5049 f	14,7220 a	0,2300 d	1,6460 e	0,1622 c
13	16,5667 e	15,2733 e	3,3621 n	18,2692 a	0,2467 d	1,3793 e	0,1933 c
14	37,1867 d	19,5333 d	4,8353 l	13,6811 b	0,5360 c	4,2213 d	0,1856 c
15	39,4667 d	26,7067 c	5,0429 k	10,4036 b	0,6840 c	6,9560 d	0,2389 b
16	48,3267 c	17,9067 d	0.0001 s	12,7381 b	0,7547 b	5,6673 d	0,2411 b
17	25,3933 e	18,5800 d	6,6239 i	16,0492 a	0,4520 c	2,8787 e	0,2056 c
18	70,4333 a	22,9800 d	0.0001 s	10,9125 b	1,4207 a	13,0807 b	0,3000 a
19	70,2667 a	24,8133 c	5,0609 k	10,3366 b	0,9480 b	9,5027 c	0,2044 c
20	19,4133 e	10,8867 f	8,1319 f	20,5326 a	0,1980 d	0,9873 e	0,1456 c
21	40,1267 d	11,4333 f	6,1745 j	15,5333 a	0,3760 d	2,4193 e	0,1522 c
22	27,2733 e	34,6800 b	5,9899 j	10,5650 b	0,6447 c	6,7727 d	0,1711 c
23	33,9200 d	20,0733 d	4,7286 l	13,6628 b	0,5540 c	4,2427 d	0,2467 b
24	58,1733 b	33,9467 b	0.0001 s	09,7567 b	1,1720 a	12,4613 b	0,2378 b
25	74,1600 a	20,0200 d	9,8264 d	12,8276 b	0,8567 b	6,6520 d	0,1511 c
26	61,0267 b	20,5200 d	6,0577 j	11,6532 b	0,6960 c	6,2127 d	0,1833 c
27	41,3533 d	10,6600 f	8,6467 f	20,9221 a	0,3393 d	1,8607 e	0,1111 c
28	42,8133 d	26,2733 c	0,0232 s	11,8041 b	0,8487 b	7,4700 d	0,2322 b
29	26,6867 e	14,0300 e	9,0014 e	17,4344 a	0,4160 d	2,4653 e	0,1744 c
30	36,2267 d	15,0667 e	8,3004 f	13,7808 b	0,3827 d	2,9047 e	0,1789 c
31	23,0933 e	12,2467 f	7,3910 h	13,2641 b	0,1750 d	1,3360 e	0,1967 c
32	35,8467 d	15,9867 e	4,7007 l	10,8509 b	0,4653 c	4,2100 d	0,2311 b
33	25,6533 e	10,3667 f	9,5325 d	17,6395 a	0,2220 d	1,2600 e	0,1456 c
34	38,7467 d	10,5800 f	7,9313 g	14,4253 a	0,2833 d	1,9353 e	0,1522 c
35	38,3133 d	42,8967 a	0.00002 s	07,0460 b	1,3760 a	19,1533 a	0,3589 a
36	47,0067 c	12,8600 e	11,3803 b	17,0406 a	0,5220 c	3,0247 e	0,1589 c
37	37,0667 d	24,2600 c	9,3519 d	11,9908 b	0,6467 c	5,8967 d	0,2344 b
38	61,5267 b	21,1600 d	0,7221 r	10,6813 b	0,8607 b	8,5020 c	0,2200 b
39	50,3133 c	37,3800 b	4,0860 m	08,2750 b	1,0473 b	13,2807 b	0,2456 b
40	56,9400 b	30,5467 c	0.0005 s	08,0580 b	0,9540 b	12,4047 b	0,2989 a
41	62,1933 b	22,7267 d	5,8199 j	13,0250 b	0,9053 b	7,6547 d	0,1733 c
42	23,7400 e	14,8933 e	2,6744 o	19,1296 a	0,4507 c	2,4040 e	0,2022 c
43	32,7800 d	32,9533 b	2,8519 o	09,2896 b	0,9280 b	11,0007 c	0,2867 a
44	76,2200 a	23,6933 c	0,6998 r	14,517 a	1,3340 a	9,3827 c	0,2000 c
45	57,5667 b	8,68670 f	3,9543 m	17,4649 a	0,3847 d	2,2540 e	0,1178 c
46	21,6733 e	31,8267 b	9,4551 d	09,5658 b	0,6480 c	6,8993 d	0,2367 b
47	48,7400 c	25,8933 c	0.0001 s	09,3713 b	1,0267 b	10,5853 c	0,2956 a
48	33,1933 d	28,7667 c	5,2098 k	10,9085 b	0,8300 b	7,5873 d	0,2256 b
49	14,1467 e	11,7000 f	10,6610 c	16,9134 a	0,1747 d	1,0220 e	0,1722 c

* Médias seguidas de mesma letra constituem grupo homogêneo pelo critério de Scott-Knott, a 5% de probabilidade

Tabela 4 – Médias de três características de qualidade de fruto em quarenta e nove acessos de *C.chinense*

Acessos	TSS (%)	Vit C (mg.100g MF^{-1})	Cor extraível (Unid ASTA.mg MS^{-1})
1	8,3653 c	98,6493 a	88,7795 e
2	8,4947 c	81,4104 b	229,0798 c
3	6,4840 c	83,6497 b	123,9669 d
4	8,9027 b	91,6471 b	105,5372 e
5	7,2053 c	86,4221 b	181,9160 d
6	8,5933 b	80,9483 b	56,6635 e
7	9,4413 b	86,3155 b	30,3525 e
8	8,2920 c	101,3151 a	75,2537 e
9	8,8000 b	106,1491 a	56,7479 e
10	8,3000 c	100,9241 a	60,2101 e
11	7,7467 c	99,4668 a	130,9284 d
12	8,2653 c	78,6024 b	234,5678 c
13	7,6413 c	118,9450 a	383,7431 b
14	8,8120 b	97,4764 a	273,7678 c
15	7,1907 c	107,5709 a	140,4897 d
16	9,0867 b	113,1513 a	38,9561 e
17	7,8760 c	112,0495 a	111,5031 e
18	7,0347 c	100,0711 a	133,8556 d
19	6,5907 c	111,6585 a	174,5107 d
20	10,8693 a	76,0077 b	34,8722 e
21	11,7480 a	119,0161 a	80,8782 e
22	6,7547 c	56,1740 c	50,6970 e
23	7,4627 c	113,4002 a	62,8214 e
24	7,2027 c	81,6237 b	54,6571 e
25	8,0920 c	134,3001 a	185,3949 d
26	7,8360 c	103,6611 a	158,2998 d
27	12,9053 a	105,4738 a	595,8413 a
28	7,4693 c	91,8959 b	188,4510 d
29	9,3067 b	115,9593 a	315,2386 b
30	7,0507 c	127,0136 a	99,6102 e
31	9,0827 b	135,2243 a	72,3165 e
32	7,8653 c	108,2463 a	171,1639 d
33	9,6880 b	97,1209 a	49,2786 e
34	8,8480 b	117,9854 a	65,9074 e
35	6,7560 c	73,9817 b	87,4950 e
36	7,9867 c	138,8498 a	109,1625 e
37	7,8000 c	110,0945 a	328,7361 b
38	7,2053 c	76,5408 b	113,1116 e
39	7,0427 c	116,0660 a	154,6405 d
40	5,9347 c	101,3151 a	214,3756 d
41	5,9720 c	70,6050 b	137,9347 d
42	8,6960 b	112,9025 a	44,5041 e
43	6,1960 c	86,8131 b	97,2475 e
44	9,1773 b	51,3756 c	152,5349 d
45	8,4667 c	38,7574 c	286,3567 c
46	7,8000 c	89,4078 b	128,4669 d
47	7,1720 c	74,9058 b	142,1571 d
48	8,2973 c	96,5522 a	81,3665 e
49	8,0427 c	107,5709 a	108,7209 e

* Médias seguidas de mesma letra constituem grupo homogêneo pelo critério de Scott-Knott, a 5% de probabilidade

As médias das Tabelas 3 e 4 foram utilizadas no cálculo da distância Euclidiana, utilizando o método UPGMA.

O número ótimo de grupos, estabelecido por meio da análise gráfica dos valores da raiz-quadrada do desvio padrão médio (RMSSTD), foi determinado geometricamente de acordo com Larson et al. (1998). O ponto que estabelece um declínio acentuado do RMSSTD nos indica no eixo horizontal o número ótimo (Figura 4). Este valor ótimo é melhor observado na Figura 5 ao se plotar as distâncias entre os pontos de uma reta que corta a curva estimada e esta própria curva, pois geometricamente a maior distância corresponde a maior curvatura. A reta que corta a curva estimada foi determinada através de procedimentos algébricos. As análises gráficas representadas por este método são mostradas nas Figuras 4 e 5.

Em relação ao R-Squared (RS), o modelo proposto pelo método da máxima curvatura não se ajustou bem a ele, uma vez que obtivemos um coeficiente de determinação (R^2) igual a 0,53, que é baixo. Porém, apresentamos na Figura 3 o gráfico do RS em relação ao número de grupos, uma vez que o fato do RS não ter se ajustado bem ao modelo proposto e aos dados deste trabalho, não quer dizer que para outro modelo e outros dados ele não se ajustará bem. Apesar disso, o ponto de máxima curvatura da curva do RS nos indica no eixo horizontal o número ótimo.

A informação da raiz-quadrada do desvio padrão médio (Figura 4) permitiu a identificação do número ótimo de grupos em relação à máxima curvatura. Com a obtenção do número ótimo, os 49 acessos foram agrupados em 7 grupos (Tabela 4).

Assim, a metodologia proposta neste trabalho determinou, então, a formação de sete agrupamentos (Tabela 5), sendo o maior grupo composto por vinte e quatro acessos e os grupos 5, 6 e 7 foram formados apenas por um acesso cada: o acesso 13 pertencente ao grupo 5, o acesso 45 pertencente ao grupo 6 e o acesso 27 pertencente ao grupo 7.

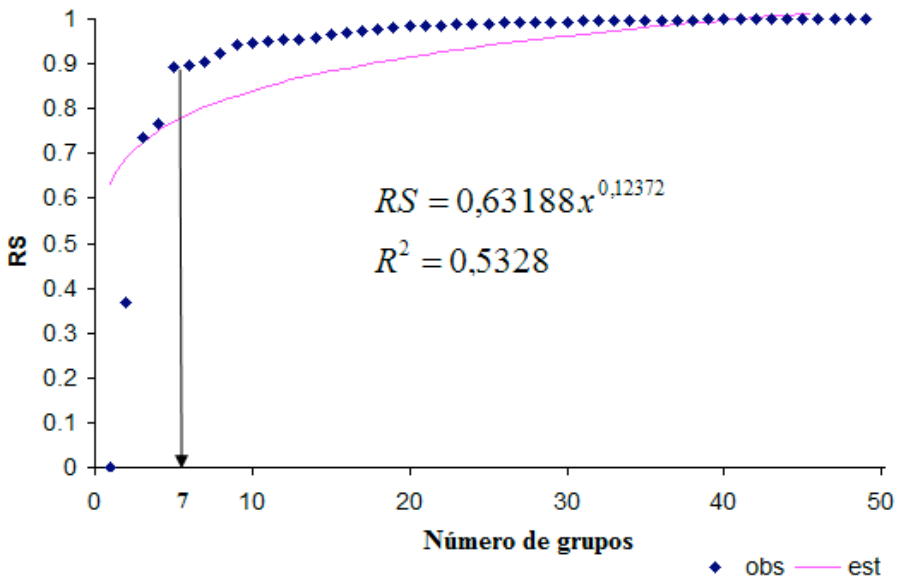


Figura 3 – Comportamento do R-Squared (RS) em função do número de grupos e o respectivo ponto de máxima curvatura geométrica fornecendo o valor ótimo (indicado pela seta).

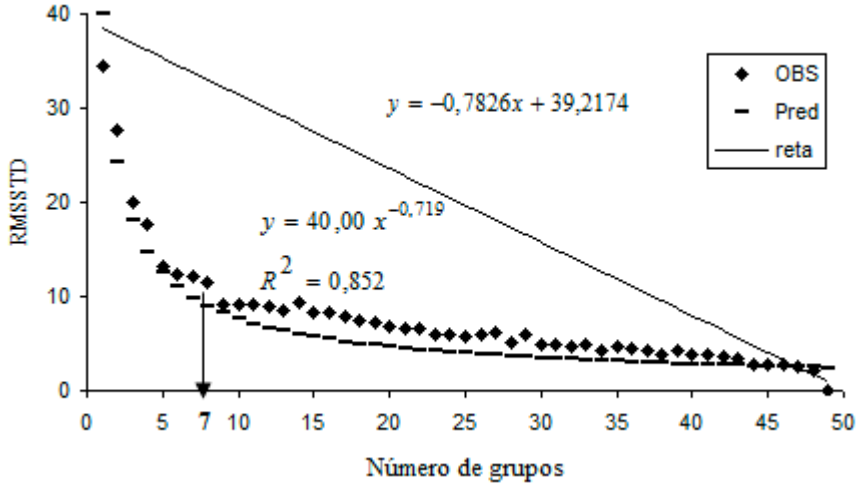


Figura 4 – Comportamento do RMSSTD em função do número de grupos e o respectivo ponto de máxima curvatura geométrica fornecendo o valor ótimo (indicado pela seta).

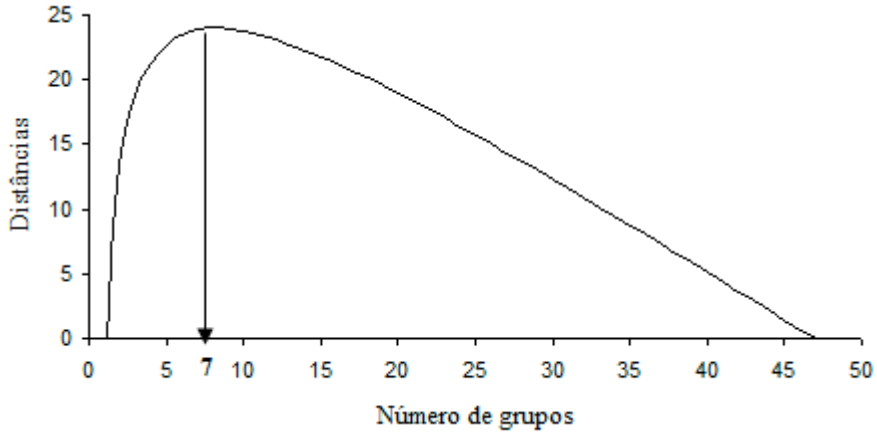


Figura 5 – Método da máxima curvatura geométrica para identificar o número de grupos.

Tabela 5 – Identificação dos acessos de *C. chinense* reunidos em 7 grupos pelo método UPGMA

Grupos	Acessos
1	5 28 3 46 11 15 41 47 19 26 32 39 38 18 25 40 44
2	2 12 14
3	9 10 1 48 23 34 17 49 33 42 21 30 36 4 43 8 31 7 20 6 24 16 35 22
4	29 37
5	13
6	45
7	27

O intervalo de confiança (95%) para os parâmetros *a* e *b* do modelo proposto pelo método da máxima curvatura modificado foi obtido no procedimento PROC NLIN (SAS, 1999) sendo [36,1969 ; 43,7942] para o parâmetro *a* e [0,6663 ; 0,7724] para o parâmetro *b*.

Na Figura 6 é apresentado o dendrograma originado com a utilização do método UPGMA e o ponto de corte obtido (número ótimo de grupos)

através do método proposto neste trabalho, utilizando-se os índices RMSSTD e RS, juntamente com o Método da Máxima Curvatura Modificado.

O acesso 27 mostrou-se bastante divergente dos demais visto que formou um grupo exclusivo e permaneceu isolado dos demais acessos no dendrograma.

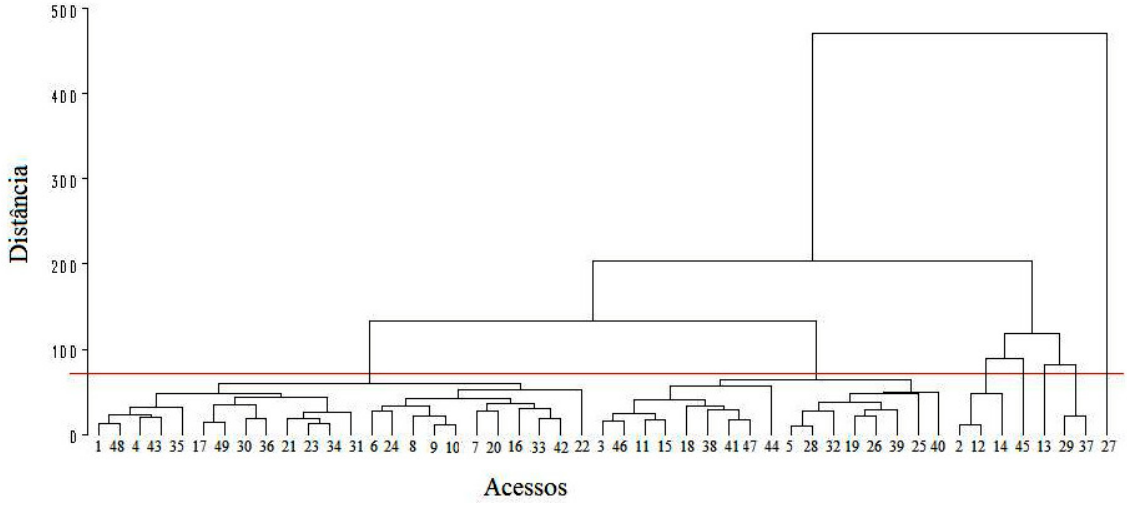


Figura 6 – Dendrograma estabelecido do padrão de dissimilaridade quantificado pelo método UPGMA, de 49 acessos de *C. chinense*, baseado na distância Euclidiana.

Baseado na distância Euclidiana e no método UPGMA, obteve-se que os acessos 5 e 28 foram os mais similares geneticamente, possuindo a menor distância (9,47). Eles correspondem, respectivamente, aos acessos BGH 1714-11 e BGH 4744-57. Entre os acessos 27 e 7 houve a maior magnitude (566,07), sendo portanto os acessos mais dissimilares.

Na Tabela 6, as médias das variáveis em cada grupo formado.

Tabela 6 – Médias das variáveis em relação a cada grupo formado com os acessos de *Capsicum Chinense*

	Médias						
Variáveis	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7
C	51.465	27.442	36.645	31.877	16.567	57.567	41.353
L	25.109	17.891	19.554	19.145	15.273	8.687	10.660
%MS	10.808	13.773	13.710	14.710	18.270	17.467	20.923
mst	0.871	0.392	0.576	0.531	0.247	0.385	0.339
peso (mft)	8.402	3.015	5.093	4.181	1.379	2.254	1.861
capTot	3.567	6.013	5.271	9.177	3.362	3.954	8.647
tss	7.283	8.524	8.431	8.553	7.641	8.467	12.905
vitC	94.539	85.830	101.308	113.027	118.945	38.757	105.474
corExt	154.835	245.805	71.814	321.987	383.743	286.357	595.841
Esp	0.230	0.174	0.205	0.204	0.193	0.118	0.111
Nº acessos	17	3	24	2	1	1	1

em que

Comp (mm): Comprimento; larg (mm): Largura; % MS: Porcentagem de matéria seca; mst (g): massa da matéria total do fruto maduro seco; peso (g): massa da matéria total do fruto maduro fresco ; capTot (mg.g MS^{-1}): Capsaicina total; tss (^{o} Brix): Teor de Sólidos solúveis; vit C (mg/100 g de fruto fresco): Vitamina C; cor ext (unid. ASTA de cor): Cor extraível; Esp (cm): espessura do pericarpo.

Note que, na Tabela 5, o Grupo 7 apresenta boa porcentagem de matéria seca (%MS), indicando que o acesso 27, único indivíduo do grupo, é um acesso indicado para a indústria, visto que os custos de secagem com ele serão reduzidos. Observa-se também um alto teor de sólidos solúveis (tss), que, comparado com os demais grupos, foi o maior. Na indústria, o alto teor de sólidos solúveis e a alta porcentagem de matéria seca são características de grande importância e presentes no acesso 27. Assim, acesso 27 (BGH 4733-56) foi considerado de maior potencial visando o cultivo sem necessidade de ser submetido ao melhoramento. Pode-se afirmar que este acesso tem alto teor

de capsaicina (capTot), de vitamina C, de sólidos solúveis, de cor extraível e da porcentagem de matéria seca, entretanto devido ao pericarpo fino (0,111 cm), o rendimento de industria pode não ser satisfatório.

Já o acesso 13 (BGH1747-26), Grupo 5, tem alto teor de vitamina C, de cor extraível e de alta porcentagem de matéria seca, e além disso, possui espessura do pericarpo (0,193 cm) maior que a do acesso 27.

Ao se comparar os resultados obtidos utilizando os índices RMSSTD e RS com o método de otimização de Tocher utilizando a Distância Euclidiana, tem-se que o método de Tocher detectou a formação de somente quatro grupos, sendo os grupos 2, 3 e 4 formados por apenas um acesso cada. Já o grupo 1 continha aproximadamente 93,88% dos acessos (Tabela 7).

Tabela 7 – Agrupamento de quarenta e nove acessos de *C. chinense* pelo método de Tocher, utilizando a distância Euclidiana e o método UPGMA

GRUPOS	ACESSOS
1	3 43 47 28 5 15 48 38 11 32 23 4 7 21 13 6 1 16 26
	9 10 14 17 2 12 46 37 30 19 41 22 42 34 8 31 44
	39 24 36 29 49 33 25 40 20 18
2	35
3	27
4	45

Por meio do Método de Tocher, pode-se verificar a dificuldade em analisar a divergência entre os acessos de C*apsicum chinense* visto que a maioria dos acessos encontravam-se agrupados numa única chave.

Na Tabela 8, os resultados fornecidos pelo aplicativo computacional GENES (CRUZ, 2001), em relação ao Método de Mojema (1977), em cada estágio do agrupamento.

Tabela 8 – Resultados obtidos pelo método UPGMA, para estimação do ponto de corte no dendrograma através do Método de Mojema (1977)

Estágio	Acesso x	Acesso y	Distância	Dist (%)	
1	5	28	9.4691	2.0425	
2	2	12	11.482	2.4767	12.2547
3	9	10	11.9058	2.5681	12.5796
4	1	48	12.5673	2.7108	13.0247
5	23	34	12.7331	2.7466	13.2688
6	17	49	14.9069	3.2155	14.3997 *
7	3	46	16.405	3.5386	15.6283 *
8	11	15	16.7166	3.6059	16.4311 *
9	41	47	16.8707	3.6391	16.9854
10	33	42	18.6443	4.0217	17.8599 *
11	21	23	18.9174	4.0806	18.5329 *
12	30	36	19.3236	4.1682	19.1125 *
13	4	43	19.5719	4.2218	19.5966
14	19	26	21.1123	4.554	20.2732 *
15	8	9	21.3764	4.611	20.8462 *
16	29	37	21.7455	4.6906	21.3632 *
17	1	4	22.9656	4.9538	21.9819 *
18	3	11	23.8924	5.1537	22.634 *
19	21	31	25.7378	5.5518	23.4671 *
20	7	20	26.8268	5.7867	24.322 *
21	5	32	27.0489	5.8346	25.0603 *
22	6	24	27.4756	5.9266	25.7381 *
23	19	39	28.3053	6.1056	26.421 *
24	38	41	29.2308	6.3052	27.1205 *
25	16	33	29.6109	6.3872	27.7669 *
26	1	35	31.9344	6.8884	28.6041 *
27	18	38	33.1531	7.1513	29.4728 *
28	6	8	33.3207	7.1874	30.2448 *
29	17	30	33.6242	7.2529	30.9529 *
30	7	16	34.8987	7.5278	31.708 *
31	5	19	37.6051	8.1116	32.6583 *
32	3	18	40.0246	8.6335	33.7541 *
33	6	21	40.12	8.6541	34.7271 *
34	6	7	42.3562	9.1365	35.8144 *
35	1	17	45.0398	9.7153	37.0494 *
36	5	25	46.6311	10.0586	38.3019 *
37	2	40	48.1867	10.3941	39.5658 *
38	1	3	50.7399	10.9449	40.9363 *
39	14	29	52.0647	11.2306	42.2829 *
40	6	22	57.1413	12.3257	43.9645 *
41	2	5	66.0143	14.2396	46.3502 *
42	1	6	70.064	15.1132	48.8566 *
43	2	44	76.0559	16.4057	51.6671 *
44	14	45	80.0387	17.2648	54.533 *
45	13	14	94.3494	20.3517	58.5144 *
46	1	2	108.7957	23.4678	63.5049 *
47	1	13	210.8883	45.4897	81.5134 *
48	1	27	463.5955	100.	**135.2742 * → θ**

NCorte = média + kDP k = 1,25

Assim, o Método de Mojema sugere um corte no dendrograma na altura de θ=135,2742. Desta forma, pode-se considerar que seria possível efetuar três cortes em diferentes estágios de agrupamento, de forma que o dendrograma poderia ser representado como na Figura 7.

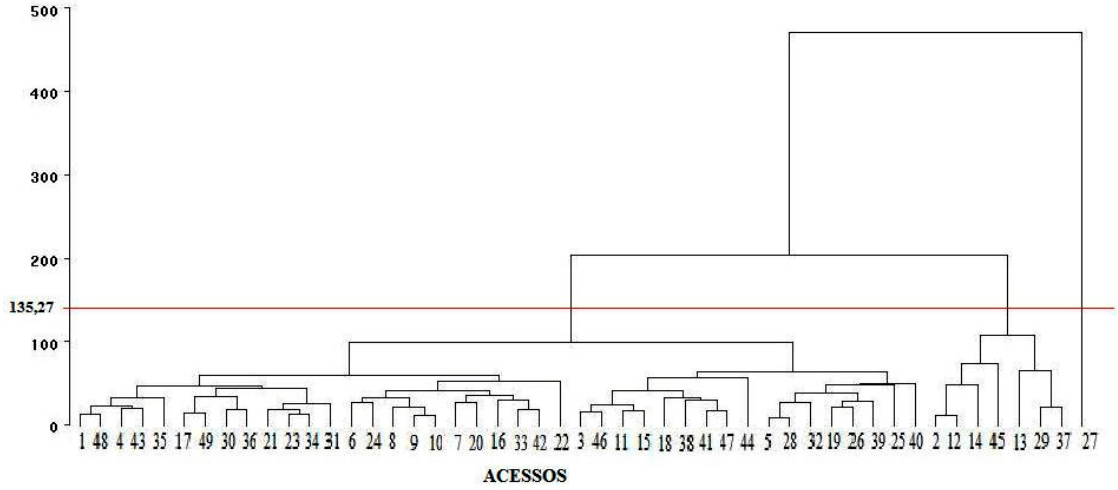


Figura 7 – Dendrograma estabelecido do padrão de dissimilaridade quantificado pelo método UPGMA, de 49 acessos de *C. chinense*, baseado na distância Euclidiana.

Observando o dendrograma, pode-se identificar os três grupos formados pelo Método de Mojema: o primeiro grupo contendo somente o acesso 27, o segundo grupo os acessos 2, 12, 14, 45, 13, 29, 37 e o terceiro grupo contendo o restante dos acessos.

Comparando o presente trabalho com o realizado por Lannes (2005), tem-se que através da distância de Mahalanobis, sua pesquisa indicou que baseado no método de otimização de Tocher, houve a formação dez grupos, como pode ser visto na Tabela 9.

Percebe-se que o número de grupos formados (oito) foi diferente do número obtido utilizando-se o método proposto neste trabalho (sete grupos), e que, apesar disso e do fato de termos usado uma medida de dissimilaridade diferente, os acessos 2 e 14 permaneceram no mesmo grupo nos dois trabalhos, o mesmo ocorrendo com os acessos 29 e 37, num outro grupo. Outra semelhança encontrada nos trabalhos é que os acessos 13, 45 e 27 permaneceram isolados dos demais, formando um único grupo cada um.

Tabela 9 – Agrupamento de quarenta e nove acessos de *C. chinense* pelo método de Tocher (Rao,1952) obtidos pela distância de Mahalanobis e pelo Método UPGMA

Grupos	Acessos
1	1 11 2 14 15 23 32 48 19 41 26 17 21 9 22 39
2	6 7 16 28 47 38 18 40 24 44 3 4 10 5
3	30 34 31 20 12 29 33 37 46
4	36 49
5	13 42 43
6	45
7	35
8	27
9	25
10	8

Na Tabela 10, um resumo do resultado obtido aplicando o Teste de Mantel entre as medidas de dissimilaridade Distância Euclidiana (DE), Distância Euclidiana Padronizada (DEP) e Distância de Mahalanobis (MAH). Para isso, foram correlacionadas as matrizes de distância de cada uma das medidas acima descritas.

Tabela 10 – Correlações entre as medidas de dissimilaridade e Teste de Mantel

Matriz de dissimilaridade	Correlação	Alfa(%)	**rcrit(1%)**			**rcrit(5%)**	
DE x DEP	0.44604**	0	-0.05419	0.33159	++	-0.04098	0.07637
DE x MAH	0.02766	65.44833	-0.05347	0.07379		-0.04066	0.06139
DEP x MAH	0.41408**	0	-0.05739	0.30058	++	-0.04194	0.08971

**,* : Significativo a 1 e 5% de probabilidade, pelo teste t.
++,+ : Significativo a 1 e 5% de probabilidade, pelo teste de Mantel baseado em 100 simulações.

Observando a Tabela 10, pelo Teste de Mantel, conclui-se que as matrizes de distância Euclidiana e a Distância Euclidiana Padronizada tiveram

correlação significativa a 1% de probabilidade. Pode-se concluir então que, a utilização de uma ou outra na análise de agrupamento realizada neste trabalho com os acessos de *Capsicum Chinense* não altera significativamente os resultados obtidos.

5. CONCLUSÃO

A proposta de estabelecimento do número ótimo de grupos em método de agrupamento hierárquico foi eficiente, e consiste em:

a) Estabelecer um método de agrupamento;
b) Obter valores de RMSSTD;
c) Obter o número de grupos utilizando o método da máxima curvatura.

A aplicação da metodologia proposta aos dados de pimenta permitiu concluir sobre a existência de sete grupos. Este número de grupos foi superior ao obtido pelo Método de otimização de Tocher (apenas quatro grupos) e ao Método de Mojema (apenas três grupos), sendo portanto de maior poder de discriminação.

6. REFERÊNCIAS

ARUNACHALAM, V. **Genetic distance in plant breeding.** The Indian Journal of Genetic and Plants Breeding, v.41, n.2, p.226-236. 1981.

BEAUMONT, M.A.; IBRAHIM, K.M.; BOURSOT, P.; BRUFORD, M.W. Measuring genetic distance. p. 315–325. *In* A. Karp et al. (ed.) **Molecular tools for screening biodiversity**. Chapman and Hall, London, 1998.

BOSLAND, P. W.; VOTAVA, E. J. **Peppers: vegetable and Spice Capsicums.** CABI Publishing. 2000. 204p.

BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.

CARVALHO, L. P. ; LANZA, M. A. ; FALIERI, J. ; SANTOS, J. W. Análise da diversidade genética entre acessos do banco ativo de germoplasma de algodão. **Pesquisa Agropecuária Brasileira**, Brasilia, v. 38, n. 10, p. 1149-1155, 2003.

COLE-RODGERS, P.; SMITH, D. W.; BOSLAND, P. W.; A novel statistical approach to analyze genetic resource evaluations using Capsicum as an example. **Crop Science.** v. 37 n. 3 p. 1000(3), 1997.

CRUZ, C.D. *Programa genes (versão Windows): aplicativo computacional em genética e estatística*. Viçosa: UFV, 2001.

CRUZ, C. D.; CARNEIRO, P. C. S.; **Modelos Biométricos aplicados ao melhoramento genético**- Volume 2. Viçosa: UFV, 2003. 585p.

CRUZ, C. D.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético.** 2 ed. rev. – Viçosa, MG: UFV, p.390, 2001.

EVERITT, B.S. 1993. **Cluster analysis**. 3rd ed. London: Heinemann Educational Books, 122 p.

JOHNSON, R.A. e WICHERN, D.W. **Applied Multivariate Statistical Analysis.** New Jersey-USA: Englewood Cliffs, 642p. 1992.

KARASAWA, M.; RODRIGUES, R.; SUDRÉ, C.P.; SILVA, M.P.; RIVA, E.M.; AMARAL JÚNIOR, A.T. Aplicação de métodos de agrupamento na quantificação da divergência genética entre acessos de tomateiro. **Horticultura Brasileira,** Brasília, v.23, n.4, p.1000-1005, out-dez 2005.

KHATTREE, R. & NAIK, D.N. 2000. **Multivariate data reduction and discrimination with SAS Software.** In: Wiley, J. & Sons (eds.). Cary, North Caroline.

LANDIM, P. M. B. **Geologia Quantitativa: Introdução à análise estatística de dados geológicos multivariados.** Rio Claro - SP, 2001.(Livro em CD-ROM).

LANNES, S.D. **Diversidade em *Capsicum chinense*: análise química, morfológica e molecular.** 2005. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 2005.

LARSON, R.; HOSTETLER, R.; EDWARDS, B. **Cálculo com Aplicações**. 5. ed. Rio de Janeiro: LTC, 1998.

LESSMAN, K. J.; ATKINS, R. E. Optimum plot size and relative efficiency of lattice designs for grain sorghum yield tests. **Crop Science**, v. 3, n. 5, p. 477-481, 1963.

MARDIA, A.K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis.** London: Academic Press, 1997. 518p.

MAXWELL, A.E. **Multivariate analysis in behavioural research**. London: Chapman & Hall, 1977.

MEIER, V. D.; LESSMAN, R. J. Estimation of optimum field plot shape and size for testing yield in Crambe abyssinica Hochst. **Crop Science,** v. 11, n. 5, p. 648-645, 1971.

METZ, J.; MONARD, M.C. **Estudo e Análise das Diversas Representações e Estruturas de Dados Utilizadas nos Algoritmos de Clustering Hierárquico.** 2006. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP.

MILLIGAN GW, COOPER MC: An examination of procedures for determining the number of clusters in a data set. ***Psychometrika*** 1985, 50:159-179.

MOHAMMADI, S.A., PRASANNA B.M. Analysis of genetic diversity in crop plants – Salient statistical tools and considerations. **Crop Science,** v.43, p.1235-1248, 2003.

MOJENA, R. Hierarquical grouping methods and stopping rules: an evaluation. The Computer Journal 20(4): 359-363,1977.

RAO, C. R.; **An advanced statistical method in biometric research.** New York, Ed. John Wiley e Sons, p.390, 1952.

SANTANA, C. M.; MALINOVSKI, J. R. Uso da análise multivariada no estudo de fatores humanos em operadores de motosserra, **Cerne**, v. 8, n. 2, p. 101–107, 2002.

SAS INSTITUTE. SAS System: SAS/STAT version 8.0 (software). Cary. 1999.

SCOTT, A. J.; KNOTT, M. A Cluster analysis method for grouping means in the analysis ofvariance. **Biometrics**, Washington, v. 30, p. 507-512, Sept. 1974.

SHARMA, S. **Applied multivariate techniquies.** NewYork: John Wiley, 1996. 493p.

SIEGMUND, K.D.; LAIRD, P.W.; LAIRDOFFRINGA, I.A. A comparison of cluster analysis methods using DNA methylation data. **Bioinformatics,** v. 20, n.12, p.1896-1904, 2004.

SMITH, H.F. **An empirical law describing heterogeneity in the yields of agricultural crops.** Journal of Agricultural Science, Canberra, v.28, p.1-23, 1938.

SNEATH, P.H.; SOKAL, R.R. **Numeric taxonomy: the principles and practice of numerical classification.** San Francisco: W.H. Freeman, 1973. 573 p.

SOUZA, A. P. Biologia molecular aplicada ao melhoramento. In*:* NASS, L. L.; VALOIS, A. C. C.; MELO, I. S.; VALADARES-INGLIS, M. C. (Eds). **Recursos Genéticos e Melhoramento Plantas.** Rondonópolis, Fundação MT, cap. 29, p. 939-966, 2001.

STEEL, R.G.D.; TORRIE, J.H. **Principles and procedures of statistics**. 2.ed. New York: Mcgraw-Hill, 1980. 633p.

SUDRÉ, C.P.; RODRIGUES, R.; RIVA, E.M.; KARASAWA, M.; AMARAL JÚNIOR, A.T. Divergência genética entre acessos de pimenta e pimentão utilizando técnicas multivariadas. **Horticultura Brasileira,** Brasília, v.23, n.1, p.22-27, jan.-mar. 2005.

VIANA, A. E. S. **Estimativas do tamanho de parcela e característica do material de plantio em experimentos com *Manihot esculenta* Crantz**. 1999. 132 f. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 1999.

YORINORI, J.T.; KIIHL, R.A. de S. **Melhoramento de plantas visando resistência a doenças.** In: Recursos Genéticos e Melhoramento Plantas. Rondonópolis, Fundação MT, cap.23, p.715-735, 2001.

APÊNDICES

APÊNDICE 1

Procedimentos utilizados no software SAS (1999) para o cálculo do RMSSTD:

```
proc glm data=cecon;
class Gen     Rep ;
model c       L       MS       MatS       P       ct       tss       VitC
corExt       Espp = Gen     Rep;
manova h=Gen     Rep;
lsmeans Gen / out=cecon1;
run;
proc print;run;

data c; set cecon1; keep _NAME_ Gen LSMEAN;  if _NAME_ ne "c"  then
delete;run;
data c;set c; rename LSMEAN = c;run;

data L; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "L" then
delete;run;
data L;set L; rename LSMEAN = L;run;

data MS; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "MS" then
delete;run;
data MS;set MS; rename LSMEAN = MS;run;

data MatS; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "MatS"
then delete;run;
data MatS;set MatS; rename LSMEAN = MatS;run;

data P; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "P" then
delete;run;
data P;set P; rename LSMEAN = P;run;

data ct; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "ct" then
delete;run;
data ct;set ct; rename LSMEAN = ct;run;

data tss; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "tss"
then delete;run;
data tss;set tss; rename LSMEAN = tss;run;

data VitC; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "VitC"
then delete;run;
data VitC;set VitC; rename LSMEAN = VitC;run;

data corExt; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne
"corExt" then delete;run;
data corExt;set corExt; rename LSMEAN = corExt;run;

data Espp; set cecon1; keep _NAME_ Gen LSMEAN; if _NAME_ ne "Espp"
then delete;run;
```

```
data Espp;set Espp; rename LSMEAN = Espp;run;

data cecon2; merge c      L      MS      MatS      P      ct
tss      VitC      corExt      Espp; by Gen;
proc print;run;

/******complemento da anova multivariada:analise de cluster******/

proc cluster data=Cecon2 method=AVERAGE  rmsstd noeigen nonorm
out=tree;
   id Gen;
   var c      L      MS      MatS      P      ct      tss      VitC
corExt      Espp;
run;
proc print data=tree;run;


proc gplot data=tree;
plot _rmsstd_*_ncl_;
run;


/*********determinaçao do numero de cluster***************/

proc model data=tree;
 _rmsstd_=a /(_ncl_)**(b);
  parms a=40 b=0.71;
 fit _rmsstd_ / outest=est outall out=all;
run;
proc print data=est;run;

data est;set est;
elev=1/((2*b) + 2);
a2=(a**2);
b2=(b**2);
b21=(2*b)+1;
b22=b+2;
nc=((a2*b2*b21)/(b22))**elev;
 run;proc print data=est;run;

/*****nc=[(((a**2)*(b**2)*(2*b + 1))/(b+2))**elev] ****/

 /*para achar intervalo de confiança*/

 proc nlin data=tree;
 parms a=40.00 b=0.71;
 model  _rmsstd_=a /(_ncl_)**(b);
 output;
run;

/*********determinaçao do numero de cluster***************/

proc tree data=tree nclusters=7 out=saidatree;
copy c      L      MS      MatS      P      ct      tss      VitC
corExt      Espp;
run;
proc sort data=saidatree; by CLUSTER;run;
proc print data=saidatree;run;

data media; set saidatree; proc sort;by cluster;run;
proc means  data=media;
```

```
 var c       L       MS       MatS       P       ct       tss       VitC
corExt       Espp;by cluster;
run;
```

Procedimentos utilizados no software SAS (1999) para o cálculo do RS:

```
proc cluster data=Cecon2 method=AVERAGE  rsq noeigen nonorm
out=tree;
   id Gen;
   var c       L       MS       MatS       P       ct       tss       VitC
corExt       Espp;
run;
proc print data=tree;run;
/*
data tree1; set tree;keep _DIST_ _NAME_;
 if _DIST_ = 0 then delete;
proc print;run;
data tree1;set tree1; rename _NAME_= CLUSNAME;proc print;run;
*/
proc gplot data=tree;
plot _rsq_*_ncl_;
run;

data tree1; set tree; keep _rsq_ _ncl_;
proc model data=tree1;
 _rsq_=a /(_ncl_)**(b);
  parms a=0.05 b=-0.001;
 fit _rsq_ / outest=est1 outall out=all1;
run;
```

APÊNDICE 2

Output do SAS quanto ao teste F realizado

The SAS System 15:39 Monday, November 10, 2008 5

The GLM Procedure

Dependent Variable: c

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	36652.13436	733.04269	12.24	<.0001
Error	96	5750.46798	59.90071		
Corrected Total	146	42402.60234			

R-Square	Coeff Var	Root MSE	c Mean
0.864384	18.81192	7.739555	41.14177

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	36550.58154	761.47045	12.71	<.0001
Rep	2	101.55282	50.77641	0.85	0.4316

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	36550.58154	761.47045	12.71	<.0001
Rep	2	101.55282	50.77641	0.85	0.4316

The SAS System 15:39 Monday, November 10, 2008 6

The GLM Procedure

Dependent Variable: L

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	9017.228577	180.344572	22.86	<.0001
Error	96	757.405351	7.889639		
Corrected Total	146	9774.633928			

R-Square	Coeff Var	Root MSE	L Mean
0.922513	13.45752	2.808850	20.87197

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	9003.991661	187.583160	23.78	<.0001
Rep	2	13.236916	6.618458	0.84	0.4353

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	9003.991661	187.583160	23.78	<.0001
Rep	2	13.236916	6.618458	0.84	0.4353

The GLM Procedure

Dependent Variable: MS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	1551.285316	31.025706	4.35	<.0001
Error	96	684.687746	7.132164		
Corrected Total	146	2235.973061			

R-Square	Coeff Var	Root MSE	MS Mean
0.693785	20.44144	2.670611	13.06469

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	1547.204528	32.233428	4.52	<.0001
Rep	2	4.080788	2.040394	0.29	0.7518

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	1547.204528	32.233428	4.52	<.0001
Rep	2	4.080788	2.040394	0.29	0.7518

The GLM Procedure

Dependent Variable: MatS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F

Model	50	14.42583086	0.28851662	6.05 <.0001
Error	96	4.57756713	0.04768299	
Corrected Total	146	19.00339798		

R-Square	Coeff Var	Root MSE	MatS Mean
0.759118	33.60542	0.218364	0.649789

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	14.40033113	0.30000690	6.29	<.0001
Rep	2	0.02549973	0.01274986	0.27	0.7659

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	14.40033113	0.30000690	6.29	<.0001
Rep	2	0.02549973	0.01274986	0.27	0.7659

The GLM Procedure

Dependent Variable: P

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	2215.545380	44.310908	14.27	<.0001
Error	96	298.133860	3.105561		
Corrected Total	146	2513.679241			

R-Square	Coeff Var	Root MSE	P Mean
0.881395	29.98436	1.762260	5.877265

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	2213.587694	46.116410	14.85	<.0001
Rep	2	1.957686	0.978843	0.32	0.7304

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	2213.587694	46.116410	14.85	<.0001
Rep	2	1.957686	0.978843	0.32	0.7304

The GLM Procedure

Dependent Variable: ct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	1882.149868	37.642997	552.49	<.0001
Error	96	6.540824	0.068134		
Corrected Total	146	1888.690692			

R-Square	Coeff Var	Root MSE	ct Mean
0.996537	5.340479	0.261024	4.887653

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	1881.774099	39.203627	575.39	<.0001
Rep	2	0.375769	0.187885	2.76	0.0685

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	1881.774099	39.203627	575.39	<.0001
Rep	2	0.375769	0.187885	2.76	0.0685

The GLM Procedure

Dependent Variable: tss

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	260.2918200	5.2058364	4.37	<.0001
Error	96	114.2727464	1.1903411		
Corrected Total	146	374.5645664			

R-Square	Coeff Var	Root MSE	tss Mean
0.694918	13.43734	1.091028	8.119374

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	259.6087478	5.4085156	4.54	<.0001
Rep	2	0.6830722	0.3415361	0.29	0.7512

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	259.6087478	5.4085156	4.54	<.0001
Rep	2	0.6830722	0.3415361	0.29	0.7512

The GLM Procedure

Dependent Variable: VitC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	63161.12009	1263.22240	4.41	<.0001
Error	96	27486.04549	286.31297		
Corrected Total	146	90647.16558			

R-Square	Coeff Var	Root MSE	VitC Mean
0.696780	17.32650	16.92079	97.65844

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	63120.70237	1315.01463	4.59	<.0001
Rep	2	40.41772	20.20886	0.07	0.9319

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	63120.70237	1315.01463	4.59	<.0001
Rep	2	40.41772	20.20886	0.07	0.9319

The GLM Procedure

Dependent Variable: corExt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	1603432.388	32068.648	11.58	<.0001
Error	96	265891.054	2769.698		
Corrected Total	146	1869323.442			

R-Square	Coeff Var	Root MSE	corExt Mean
0.857761	36.82341	52.62792	142.9198

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	1593510.828	33198.142	11.99	<.0001
Rep	2	9921.560	4960.780	1.79	0.1723

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	1593510.828	33198.142	11.99	<.0001
Rep	2	9921.560	4960.780	1.79	0.1723

The GLM Procedure

Dependent Variable: Espp

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	50	0.36560877	0.00731218	5.61	<.0001
Error	96	0.12508934	0.00130301		
Corrected Total	146	0.49069811			

R-Square	Coeff Var	Root MSE	Espp Mean
0.745079	17.38251	0.036097	0.207664

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gen	48	0.36513515	0.00760698	5.84	<.0001
Rep	2	0.00047362	0.00023681	0.18	0.8341

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gen	48	0.36513515	0.00760698	5.84	<.0001
Rep	2	0.00047362	0.00023681	0.18	0.8341

The GLM Procedure
Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for Gen
E = Error SSCP Matrix

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Gen Effect
H = Type III SSCP Matrix for Gen
E = Error SSCP Matrix

S=10 M=18.5 N=42.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00000003	9.00	480	883.09	<.0001
Pillai's Trace	6.73374456	4.12	480	960	<.0001
Hotelling-Lawley Trace	366.26416682	65.07	480	629.89	<.0001
Roy's Greatest Root	331.16514888	662.33	48	96	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper boun.

The GLM Procedure
Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall Rep Effect
H = Type III SSCP Matrix for Rep
E = Error SSCP Matrix

S=2 M=3.5 N=42.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.84430288	0.77	20	174	0.7487
Pillai's Trace	0.16185764	0.77	20	176	0.7411
Hotelling-Lawley Trace	0.17711251	0.76	20	143.34	0.7531
Roy's Greatest Root	0.11191504	0.98	10	88	0.4626

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.